EXHIBIT I

CABIOS

Vol. 11 no. 4 1995
Pages 435–439

SUBIM: a program for analysing the Kabat database and determining the variability subgroup of a new immunoglobulin sequence

S.Déret[2], C.Maissiat, P.Aucouturier and J.Chomilier[1]

Abstract

Although various programs are available to extract all the information included in protein sequence databases, none is dedicated to immunoglobulins. For this purpose, we designed a program, SUBIM, which is adapted to the Kabat database specialized in immunoglobulin sequences. Besides all the possibilities of any database searching program, SUBIM analyses new sequences of variable regions and determines the variability subgroup they belong to. It also numbers the new sequence according to the system established by Kabat and co-workers for an easier comparison with the other immunoglobulins, thus realizing an automatic alignment with other members of a given type of immunoglobulin chain. This program is largely machine independent and requires very little memory, and should help biochemists concerned with new immunoglobulin sequences.

Introduction

The current accumulation of immunoglobulin (Ig) sequences, and the need for comparison studies in correlation with Ig binding specificities or pathogenic properties, require the use of an exhaustive and performing dedicated database. The Kabat database (Kabat et *al.*, 1991) is presently the most complete collection of amino acid sequences of Ig and related proteins of immunological interest. The 5.0 release of Kabat database contains 469 human heavy, kappa and lambda chain variable (V) regions (V_H, V_κ and V_λ respectively) compared to the 132 found in SwissProt release 29.0 (Bairoch and Boeckman, 1991) by retrieving all files containing the keywords immunoglobulin' and human' with the Genetics Computer Group (GCG) method (Devereux *et al.*,1984). All human Ig sequences from SwissProt are included in the Kabat database, but the V_H, V_κ and V_λ of SwissProt correspond only to 26%, 29% and 33% respectively of those collected by Kabat and co-workers when taking into account sequences of 35 residues or more. In addition, many small pieces of sequences are included in the Kabat database. These are of interest because the original numbering allows all V regions of a given type (κ, λ or H) to be directly aligned, so their position can be located within the full sequence. This numbering takes into account the frequent amino acid deletions and insertions which occur at certain positions as a result of genetic divergence of the V gene segments and junctionnal variability during the somatic rearrangements of Ig genes.

To our knowledge, no dedicated software has been developed to use efficiently the Kabat database. In this paper, we present the program SUBIM which allows selection and extraction of V and constant (C) region amino acid sequences for comparison studies. It also allows these sequences to be numbered according to Kabat methodology. Compared to other available programs, SUBIM clusters Ig V regions into sets of related sequences which are called variability subgroups. Since the early studies of Milstein (1967) and Hilschmann (1969), subgroup classification has become (mainly by the monumentous work of Kabat and colleagues) a guiding principle of structural immunology. This taxonomy is based upon the conservation of residues between pairs of sequences and within one subgroup. For example, the mean sequence identities within subgroups κII and λVI are close to 80% in both cases, with a root mean square deviation of ~5, while the identity between their consensus is 44%. The identity between the consensus sequences of the four κ subgroups is between 61 and 74%. For subgroups I–IV of human α chains, this classification was fully confirmed when the human germline sequences on one allele became known (Zachau and Schäble, 1993).

Algorithm

Extraction of information concerning one or a set of sequences

A first use of SUBIM is for scanning the database, using keywords that can be selected by any combination based on Boolean algebra.

Allowed keywords are:
protein name (full or partial)
species
author
year of publication

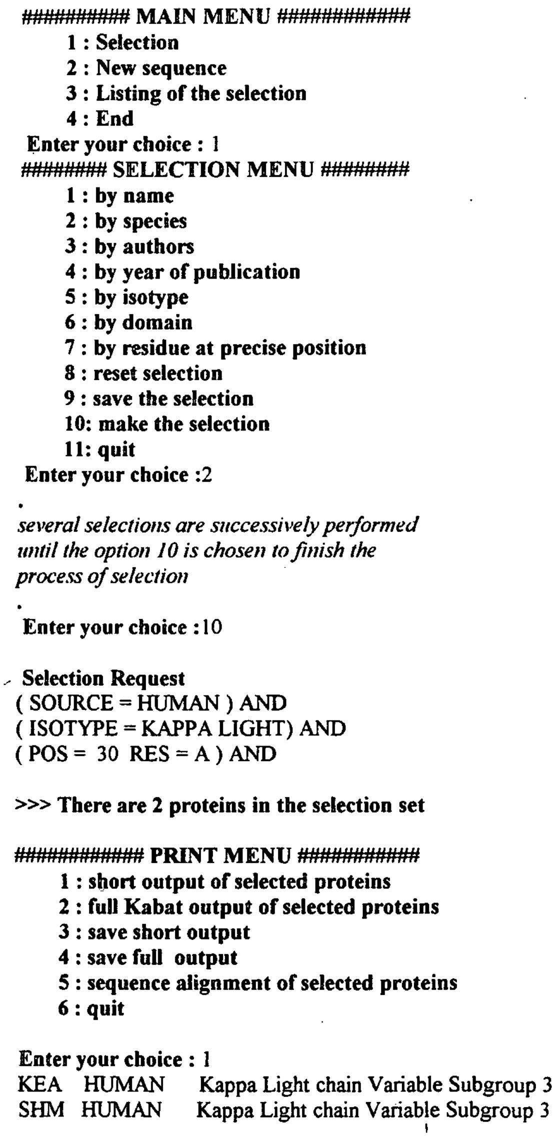
Laboratoire d'Immunologie et Immunopathologie, CNRS URA 1172, F-86021 Poitiers and [1]Laboratoire de Minéralogie-Cristallographie, Universités Paris VI et VII, CNRS URA 009, Case 115, F-75252 Paris Cédex 05, France

[2]*To whom correspondence should be addressed. Email: criso@zeus.univ-poitiers.fr*

isotype and variability subgroup
domain.

All this information can be retrieved from the database as long as these fields are documented in the files. Another possibility is retrieving information within sequences, such as selecting all Ig sequences with a given residue at a determined position.

Different output formats for the retrieval of a sequence are possible. An example of a request with keywords through SUBIM is presented on Figure 1, with the short output option. This short output only gives the protein name, species, isotype and variability subgroup. If this reduced information is insufficient, a longer output is possible, where all the information available in the database is retrieved, including the sequence and all comments. An example of a long output is shown in Figure 2.



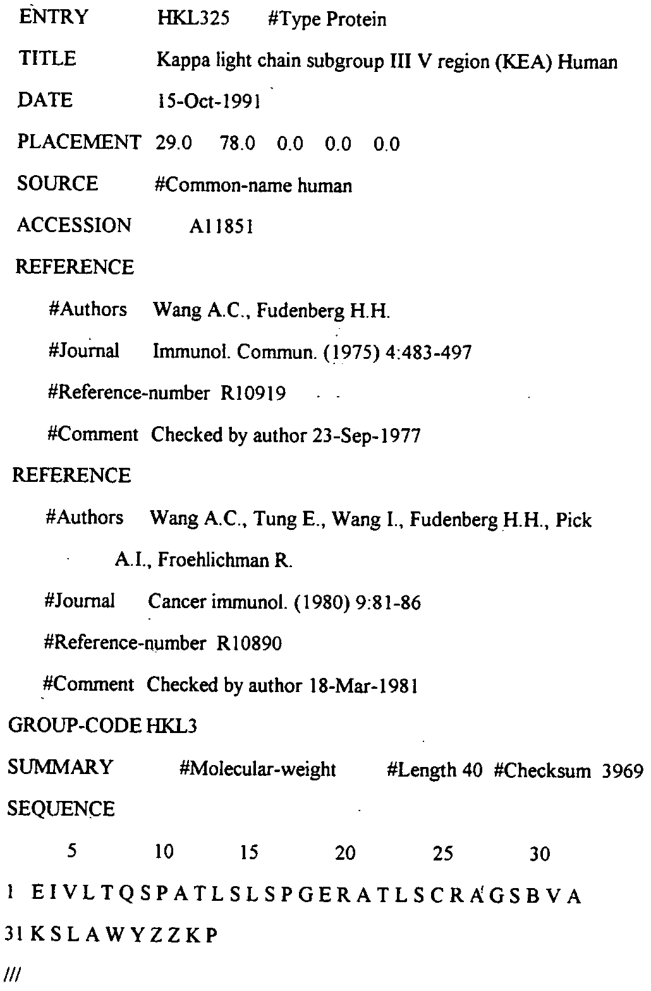
```
########## MAIN MENU ############
        1 : Selection
        2 : New sequence
        3 : Listing of the selection
        4 : End
 Enter your choice : 1
######## SELECTION MENU ########
        1 : by name
        2 : by species
        3 : by authors
        4 : by year of publication
        5 : by isotype
        6 : by domain
        7 : by residue at precise position
        8 : reset selection
        9 : save the selection
        10: make the selection
        11: quit
 Enter your choice :2

 several selections are successively performed
 until the option 10 is chosen to finish the
 process of selection

 Enter your choice :10

 Selection Request
 ( SOURCE = HUMAN ) AND
 ( ISOTYPE = KAPPA LIGHT) AND
 ( POS =  30  RES = A ) AND

>>> There are 2 proteins in the selection set

############ PRINT MENU ###########
        1 : short output of selected proteins
        2 : full Kabat output of selected proteins
        3 : save short output
        4 : save full  output
        5 : sequence alignment of selected proteins
        6 : quit

Enter your choice : 1
KEA  HUMAN    Kappa Light chain Variable Subgroup 3
SHM  HUMAN    Kappa Light chain Variable Subgroup 3
```

Fig. 1. Systematic research. Monoclonal Ig light chains of the $V_{\kappa I}$ subgroup involved in Fanconi's syndrome have been sequenced. Two of them (CHEB and TRE) include one residue that has never been described before in this subgroup: Ala at position 30. Thus, it is worth looking for this residue in every monoclonal light chain which might cause Fanconi's syndrome. This figure shows the succession of menus and the short output option. Ala 30 already exists in two human κ chains of the $V_{\kappa III}$ subgroup, KEA and SHM.

Determination of the variability subgroup of a new sequence

Because of the high homology within one subgroup, especially in the N-terminal part, the first 15 residues are sufficient to determine the variability subgroup. The N terminus is a rather accessible data even when small amounts of the purified protein are available, while the complete amino acid sequence may remain undetermined in many instances.

Consensus sequences within each subgroup were previously determined by Kabat *et al.* (1991). Comparisons



```
ENTRY          HKL325      #Type Protein
TITLE          Kappa light chain subgroup III V region (KEA) Human
DATE           15-Oct-1991
PLACEMENT      29.0    78.0    0.0    0.0    0.0
SOURCE         #Common-name human
ACCESSION        A11851
REFERENCE
   #Authors    Wang A.C., Fudenberg H.H.
   #Journal    Immunol. Commun. (1975) 4:483-497
   #Reference-number  R10919
   #Comment    Checked by author 23-Sep-1977
REFERENCE
   #Authors    Wang A.C., Tung E., Wang I., Fudenberg H.H., Pick
                  A.I., Froehlichman R.
   #Journal    Cancer immunol. (1980) 9:81-86
   #Reference-number  R10890
   #Comment    Checked by author 18-Mar-1981
GROUP-CODE HKL3
SUMMARY        #Molecular-weight       #Length 40  #Checksum  3969
SEQUENCE
      5        10        15        20        25        30
1  E I V L T Q S P A T L S L S P G E R A T L S C R A G S B V A
31 K S L A W Y Z Z K P
///
```

Fig. 2. Example of long output format. The long output option of the print menu allows the user to retrieve all information available in the database on both human κ chains of the $V_{\kappa III}$ subgroup, described in Figure 1. The figure only gives the example of KEA.

with two consensus sequences of each subgroup (4 κ and 6 λ subgroups) proved sufficient in all studied cases. The first consensus corresponds to the most common residue at each position, and the second consensus corresponds to the second highest occurrence at each position. A frequency value (F) is attributed to each residue of the consensus sequences:

$$F = \text{occurence of this residue at this position/}$$
$$\text{number of sequences in the subgroup}$$

A residue with a frequency of between 0.95 and 1.0 is considered 'invariant'.

The 15 N-terminal residues of a new sequence are compared with the $2N$ consensus, N being the number of subgroups. If the first amino acid matches that in one consensus, a S score attached to that position is given the frequency of the corresponding residue in this consensus. At the second position an equivalent value is added to this S score, and so on until the 15th residue. The result is divided by the sum of first consensus frequencies, leading to a $S0$ score.

```
query sequence:
    D  I  V  M  T  Q  S  P  E  S  L  A  V  S  L

first consensus sequence kappa I (most frequent):
    X  D  I  Q  M  T  Q  S  P  S  S  L  S  A  S  V  G  D  V  R  T

F at each position:
0.016  0.748  0.759  0.710  0.721  0.819  0.803  0.819  0.814  0.776  0.579
0.841  0.879  0.672  0.798  0.699  0.819  0.639  0.743  0.683  0.803
Σ F first consensus kappa = 15.14

second consensus sequence kappa I (second most frequent):
    Z  B  V  Z  L  M  Z  A  A  T  T  V  P  L  T  P  R  E  S  A  I

F at each position:
0.005  0.022  0.016  0.022  0.087  0.005  0.038  0.005  0.005  0.011  0.289
0.033  0.022  0.093  0.022  0.120  0.005  0.109  0.027  0.114  0.032

S0 = 0 + 0 + 0.016 + 0 + 0 + 0 + 0 + 0 + 0 + 0.776 + 0 + 0 + 0 +0 +0 / 15.14 = 0.051.
S1 = 0.748 +0.759 +0 + 0.721 + 0.819 + 0.803 + 0.819 + 0.814 + 0 + 0.579 + 0.841
      + 0 + 0 + 0.798 + 0 / 15.14 = 0.508.
S2 = 0 + 0 + 0 + 0.005 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0.672 + 0 + 0 + 0 / 15.14 = 0.049.
S3 = 0 + 0 + 0 + 0 + 0 + 0 + 0.776 + 0 + 0 + 0.672 + 0 + 0 + 0.699 + 0 + 0 / 15.14
    = 0.141.
S4 = 0 + 0 + 0 + 0 + 0 + 0 + 0.579 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 / 15.14 = 0.038.
S5 = 0 + 0 + 0 + 0 + 0.011 + 0 + 0 + 0.093 + 0 + 0.798 + 0 + 0 + 0 + 0.027 + 0 / 15.14
    = 0.061.
S6 = 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0.289 + 0.879 + 0 + 0 + 0 + 0 + 0 + 0.743
      + 0 + 0 / 15.14 = 0.126.

Smax (for kappa I) = 50.8%
Smax (for kappa II) = 50.7%
Smax (for kappa III) = 41.8%
Smax (for kappa IV) = 73.1%
```

Fig. 3. Determination of subgroup variability of a new sequence. A new N-terminal sequence of a light chain responsible for non-amyloid light-chain deposition disease was determined. For each consensus of the subgroup $V_{\kappa I}$, the frequency F of every 15 N-terminal residue is given, which is the probability of finding such a residue at this position in the subgroup. X indicates that several amino acids were found at this position, with none being found significantly more frequently than others. SUBIM compares this sequence with each consensus sequence of the κI variability subgroup. The $S0$ score is determined by summing the 15 values of F when a match is observed between the query and one of both consensus. If no match is observed, no value is added. The query sequence is shifted by 1, hence residue D becomes residue 2, and the $S1$ score is calculated. Seven scores are calculated by successive shifts. This operation is repeated for each variability subgroup. The sequence is finally assigned to the κ light-chain subgroup IV, which yields the highest score (73.1%).

In a second step, the query sequence is shifted by 1, and hence starts at position 2. A new score is then determined similarly. Six successive shifts are performed leading to scores $S0$–$S6$. The limit of six residue shifts is based on the fact that an Ig sequence may be truncated by at most six amino acids at the N-terminal part, when aligned with the family. This value is acertained from the fact that the starting of numbering, within any subgroup of the Kabat database, is included in the range −6, +6, whatever sequence is considered. This was verified for 98.7% of the sequences.

Finally seven scores are determined for each subgroup and the highest value allows a subgroup to be attributed to the new sequence.

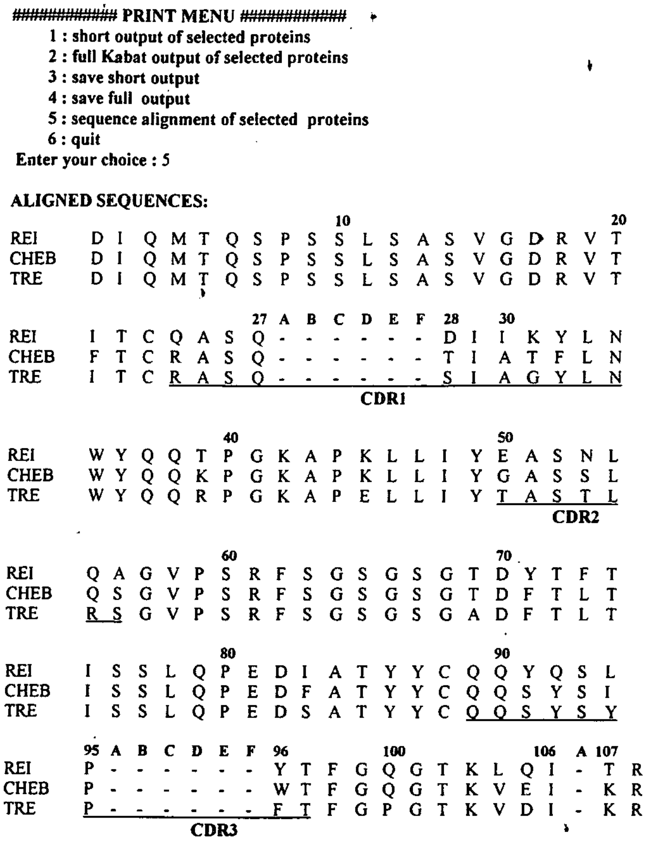
Numbering new sequences of V regions

In the Kabat database, the V region sequences are aligned to permit various comparisons of the V_{κ} and V_{λ} from different species (human, mouse, etc.) and different variability subgroups. This was performed by sequential numbering with gaps inserted for alignment. Most of these gaps are located in the complementarity determining regions (CDRs). Tramontano and Lesk (1992) demonstrated that the Kabat numbering is required for comparing Ig chains.

Strictly conserved residues within one isotype (κ, λ or heavy chain) may be used as anchor points since they are particulary important for the structure and functions. Such highly conserved residues are encountered at several positions spread all over the sequences and in a cluster at the beginning of the third framework (FR3) segment.

First, SUBIM searches for known invariant residues (e.g. Cys23, Trp35, Gly57, Cys88 in a κ chain) in the new sequence and places them at the corresponding positions. It then calculates the number of amino acids between these conserved residues. According to this difference value, gaps are filled if necessary, according to the flexible Kabat notation. For instance, if the residues numbered 27 and 28 in Kabat numbering are also found in the new sequence but separated by two residues, a gap is introduced with numbers 27A and 27B. This format procedure allows alignment of new determined amino acid sequences with sequences already available in the Kabat database. It was not developed for C domains because their sequences are highly conserved.

Figure 4 shows an example of new Ig, whose sequences are not yet in the Kabat database; the subgroup of each



```
############ PRINT MENU ############
      1 : short output of selected proteins
      2 : full Kabat output of selected proteins
      3 : save short output
      4 : save full  output
      5 : sequence alignment of selected  proteins
      6 : quit
 Enter your choice : 5

ALIGNED SEQUENCES:
                                   10                            20
REI     D  I  Q  M  T  Q  S  P  S  S  L  S  A  S  V  G  D  R  V  T
CHEB    D  I  Q  M  T  Q  S  P  S  S  L  S  A  S  V  G  D  R  V  T
TRE     D  I  Q  M  T  Q  S  P  S  S  L  S  A  S  V  G  D  R  V  T

                          27 A  B  C  D  E  F  28    30
REI     I  T  C  Q  A  S  Q  -  -  -  -  -  -  D  I  I  K  Y  L  N
CHEB    F  T  C  R  A  S  Q  -  -  -  -  -  -  T  I  A  T  F  L  N
TRE     I  T  C  R  A  S  Q  -  -  -  -  -  -  S  I  A  G  Y  L  N
                                    CDR1

                       40                            50
REI     W  Y  Q  Q  T  P  G  K  A  P  K  L  L  I  Y  E  A  S  N  L
CHEB    W  Y  Q  Q  K  P  G  K  A  P  K  L  L  I  Y  G  A  S  S  L
TRE     W  Y  Q  Q  R  P  G  K  A  P  E  L  L  I  Y  T  A  S  T  L
                                                         CDR2

                       60                            70
REI     Q  A  G  V  P  S  R  F  S  G  S  G  S  G  T  D  Y  T  F  T
CHEB    Q  S  G  V  P  S  R  F  S  G  S  G  S  G  T  D  F  T  L  T
TRE     R  S  G  V  P  S  R  F  S  G  S  G  S  G  A  D  F  T  L  T

                       80                            90
REI     I  S  S  L  Q  P  E  D  I  A  T  Y  Y  C  Q  Q  Y  Q  S  L
CHEB    I  S  S  L  Q  P  E  D  F  A  T  Y  Y  C  Q  Q  S  Y  S  I
TRE     I  S  S  L  Q  P  E  D  S  A  T  Y  Y  C  Q  Q  S  Y  S  Y

        95 A  B  C  D  E  F  96          100               106 A 107
REI     P  -  -  -  -  -  -  Y  T  F  G  Q  G  T  K  L  Q  I  -  T  R
CHEB    P  -  -  -  -  -  -  W  T  F  G  Q  G  T  K  V  E  I  -  K  R
TRE     P  -  -  -  -  -  -  F  T  F  G  P  G  T  K  V  D  I  -  K  R
                  CDR3
```

Fig. 4. Example of sequence alignment. TRE and CHEB, two $V_{\kappa I}$ Ig light chains involved in Fanconi's syndrome, are aligned with REI, a V_κ dimer whose three-dimensional structure is known. The sequence identity is 77% between TRE and REI, and 82% between CHEB and REI.

sequence has been determined and the numbering performed, allowing alignment with REI, a V_κ light chain dimer whose three-dimensional structure is available in the Protein Data Bank (Bernstein *et al.*, 1977). This alignment is a first step in the procedure of building a three-dimensional model by homology modelling.

Implementation

SUBIM has been written in the C programming language and has been implemented under the Unix operating system. The program source code is available by anonymous ftp at the node ftp.lmcp.jussieu.fr, under the pub directory. Memory requirements for the microcomputer are modest, but a hard disk with sufficient capacity to hold the Kabat database is necessary.

Discussion

We have developed a tool for exploiting the currently most complete database for Ig sequences at the protein level. SRS (Etzold *et al.*, 1993), an information indexing and retrieval system designed for libraries such as the EMBL nucleotide sequence databank (Stoehr and Cameron, 1991), the SwissProt protein (Bairoch and Boeckman, 1991) and the PROSITE library of protein subsequence consensus patterns (Bairoch, 1991) may also perform most SUBIM functions. SRS allows exploitation of all information provided by these libraries (Etzold and Argos, 1993a,b), although some knowledge of the libraries is essential for its use. Elgavish and Schroeder (1993) recently developed a program, SAW, designed for the analysis of Ig V region nucleotide sequences; it is especially interesting for comparing complementary DNA sequences extracted from GenBank or standard text files, assigning functionnal Ig V regions to known germline segments, translation to amino acid sequences and for graphical presentation of data. Our program is dedicated to the exploitation of the Ig amino acid sequences, and is thus complementary to SAW.

SUBIM was designed not only for searches but also for some useful functions, such as V region variability subgroup determination, identification of unusual residues, and numbering according to Kabat *et al.* (1991). SUBIM allows the user to extract and study known human heavy- and light-chain sequences. For instance, comparisons of primary structures of monoclonal Ig chains may be particulary relevant to the understanding of their pathogenicity (Preud'homme *et al.*, 1994). It may be also interesting to look for precise residues at certain positions, which may be related to an unusual behavior of an Ig chain: for instance, an asparagine at position 70 of a κ chain generally creates a functionnal N-glycosylation site that may cause myeloma complications such as light-chain deposition disease (Cogné *et al.*, 1991) or amyloidosis (Aucouturier *et al.*, 1992). Determining the variability subgroups may be useful because some subgroups are more frequently expressed in some complications of light-chain myeloma (Solomon *et al.*, 1982; Denoroy *et al.*, 1994).

At this time, procedures for determining variability subgroups of new sequences and formatting them, run only for human sequences but extension to other species is currently in progress.

In the last version of the Kabat database (1994), sequences are clustered by families, based on amino acid sequence only. Each family is composed of sequences that differ by 12 amino acids or less. Thus, in addition to other uses, SUBIM may be valuable in preserving the subgroup clustering.

Acknowledgements

We thank Alain Roussel for helpful suggestions. This work was supported by grants from the Association pour la Recherche contre le Cancer (6044) and INSERM (CRE 930602). S.D. is recipient of a fellowship from the Ligue Nationale contre le Cancer. This work is dedicated to the memory of our friend Claude Maissiat.

References

Aucouturier,P., Khamlichi,A.A., Preud'homme,J.L., Bauwens,M., Touchard,G. and Cogné,M. (1992) Complementary DNA sequence of human amyloidogenic immunoglobulin light-chain precursors. *Biochem. J.*, **285**, 149–152.

Bairoch,A. and Boeckman,B. (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, **19**, 2247–2249.

Bairoch,A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.*,**19**, 2241–2245.

Berstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Cogné,M., Preud'homme,J.L., Bauwens,M., Touchard,G. and Aucouturier,P. (1991) Structure of a monoclonal kappa chain of the $V_{\kappa IV}$ subgroup in the kidney and plasma cells in light chain deposition disease. *J. Clin. Invest.*, **87**, 2186–2190.

Denoroy,L., Déret,S. and Aucouturier,P. (1994) Overrepresentation of the $V_{\kappa IV}$ subgroup in light chain deposition disease. *Immunol. Lett.*, **42**, 63–66.

Devereux,J., Haeberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.

Elgavish,R.A. and Schroeder,H.W. (1993) SAW: A graphical user interface for the analysis of immunoglobulin variable domain sequences. *Biotechniques*, **15**, 1066–1071.

Etzold,T. and Argos,P. (1993a) SRS—an indexing and retrieval tool for flat file data libraries. *Comput. Applic. Biosci.*, **9**, 49–57.

Etzold,T. and Argos,P. (1993b) Transforming a set of biological flat file libraries to a fast access network. *Comput. Applic. Biosci.*, **9**, 59–64.

Hilschmann,N. (1969) Die molekularen grundlagen der antikörperbildung. *Naturwissenschaften*, **56**, 195–205.

Kabat,E.A., Wu,T.T., Perry,H.M., Gottesman,K.S. and Foeller,C. (1991) In *Sequences of Proteins of Immunological Interest*. US Department of Health and Human Services, NIH Publications, Bethesda, MD.

Milstein,C. (1967) Linked groups of residues in immunoglobulins κ chains. *Nature*, **216**, 330–332.

Preud'homme,J.L., Aucouturier,P., Touchard,G., Striker,L., Khamlichi,A.A., Rocca., A., Denoroy,L. and Cogné,M. (1994) Monoclonal immunoglobulin deposition disease (Randall type). Relationship with structural abnormalities of immunoglobulin chains. *Kidney Int.*, **46**, 965–972.

Solomon,A., Frangione,B. and Franklin,E.C. (1982) Bence Jones proteins and light chains of immunoglobulins. Preferential association of the $V_{\lambda VI}$ subgroup of human light chains with amyloidosis AL (λ). *J. Clin. Invest.*, **70**, 453–460.

Stoehr,P.J. and Cameron,G.N. (1991) The EMBL data library. *Nucleic Acids Res.*, *19*, 2227–2230.

Tramontano,A. and Lesk,A.M. (1992) Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. *Proteins*, **13**, 231–245.

Zachau,HG. and Schäble,K.F. (1993) The variable genes of the human immunoglobulin κ locus. *Biol. Chem. Hoppe-Seyler*, **374**, 1001–1022.

Received on February 20, 1995; accepted on May 19, 1995

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

□ **BLACK BORDERS**

□ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**

□ **FADED TEXT OR DRAWING**

□ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**

□ **SKEWED/SLANTED IMAGES**

□ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**

□ **GRAY SCALE DOCUMENTS**

□ **LINES OR MARKS ON ORIGINAL DOCUMENT**

☑ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**

□ **OTHER:** ____________________

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.